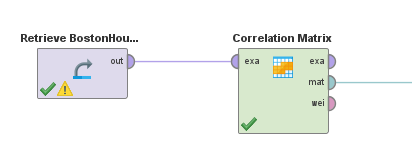
The file Boston housing contains data on housing prices, the possible goal is to create a predictive model to predict house prices:

1. Identify categorical variable(s)
2. Produce scatterplot matrix plots and commenting on the relationships among the variable you found interesting
3. Create a predictive model to predict the housing price (possible steps including data transformation, data split, model validation etc.)

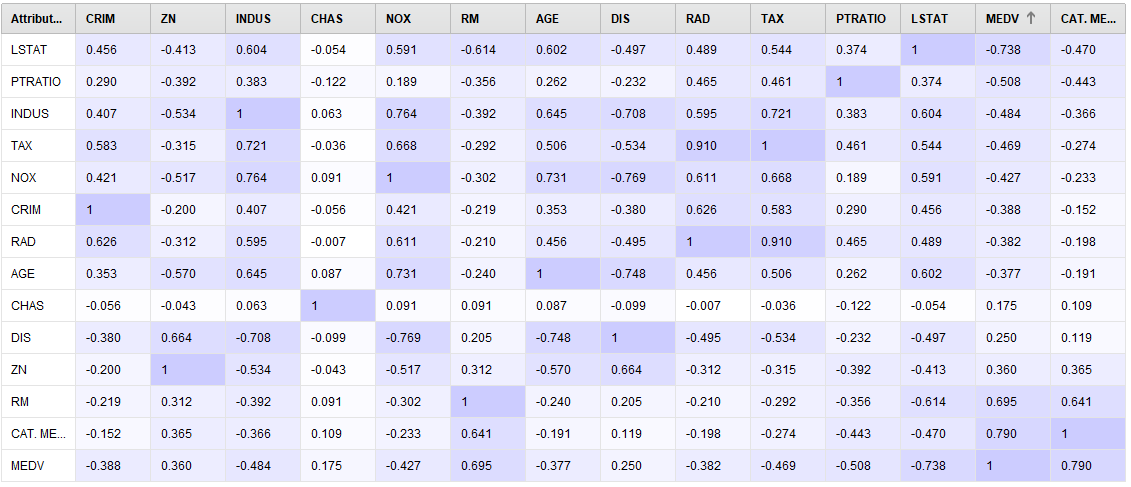
*Steps to complete this exercise:*

1. CHAS is a categorical variable with two possible outcomes: bounds river = 1, otherwise = 0

We start by running the correlation matrix



The results are shown below:



The special attribute (dependent variable) is MEDV.

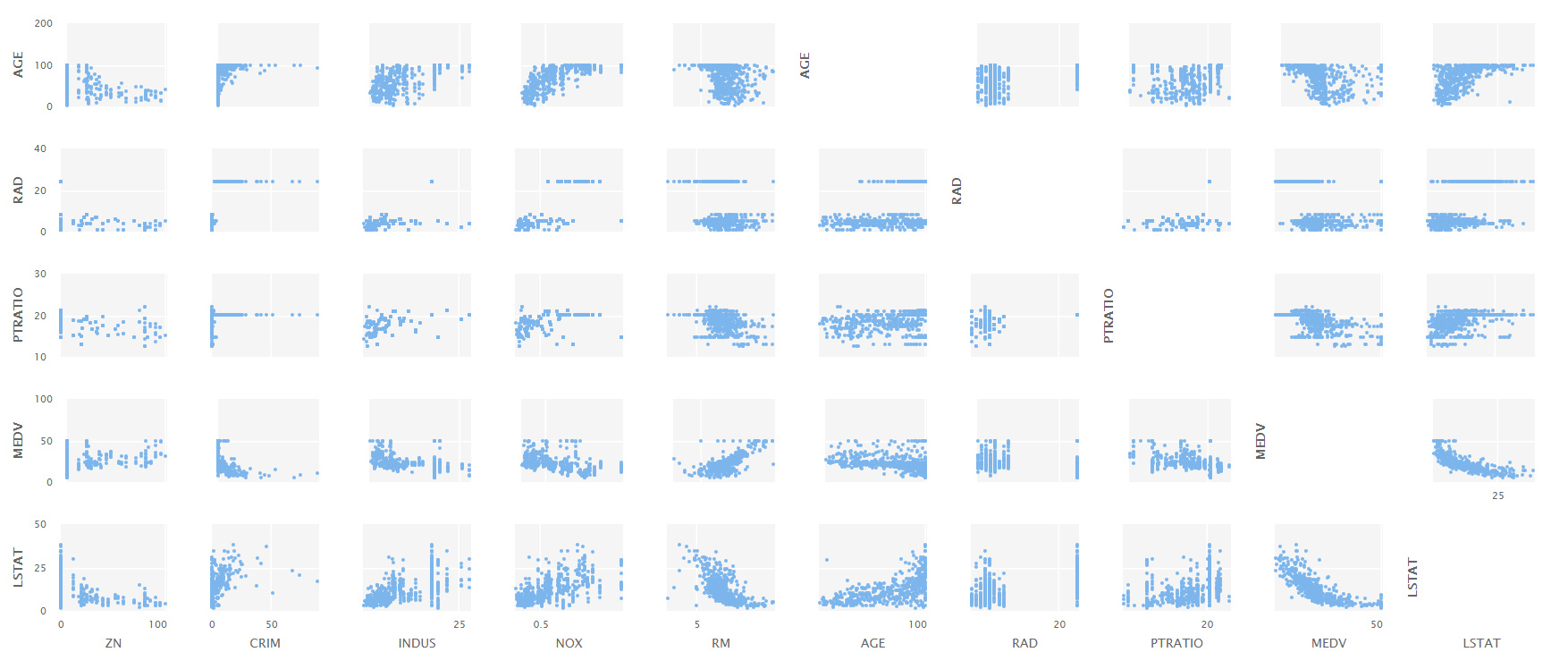
From the correlation table above, we are interested in keeping those attributes (independent variables) are highly correlated with the special attribute (MEDV) but not with each other.

DIS has low correlation with MEDV (0.25, circled in the chart), it might not be a good predictor for MEDV

LSTAT has a high negative correlation with MEDV (-0.738), it might be a good predictor for MEDV

**The general rule is to only keep attributes correlates with label (dependent variable), among the attributes, omit attributes that are highly correlated with each other to avoid multicollinearity .**

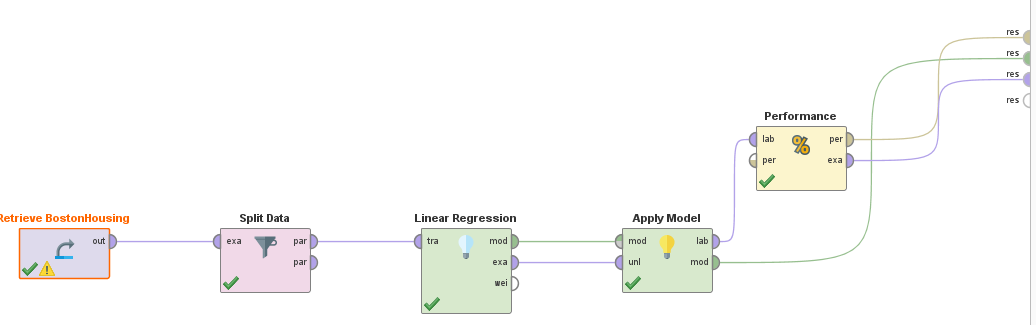
After examination of these correlations, we can generate a scatter plot matrix to visually examine the relationships.



3.

To build our model:

1st try with all possible attributes:

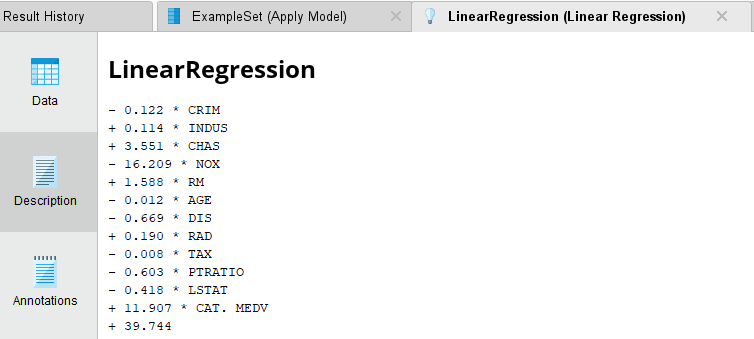


Save the results.

This set of operators generate a model based on **the training dataset (**the first split)

We can examine the results of the model, there are a few things to note:

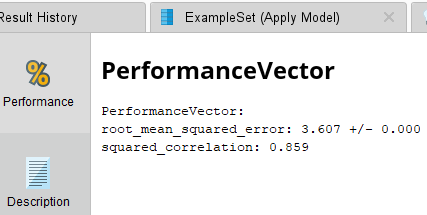
First, this is your linear regression model



To write out the equation:

Y(MEDV) = 39.744 + 11.907\*CAT.MEDV-0.418\*LSTAT-……. -0.122\*CRIM

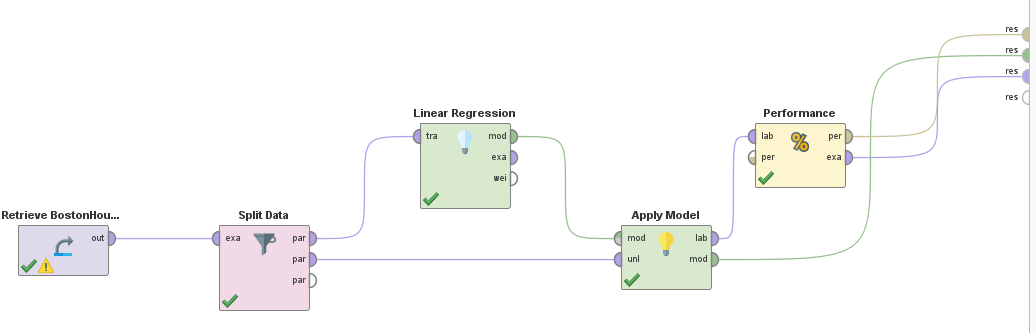
Second, we need to look at



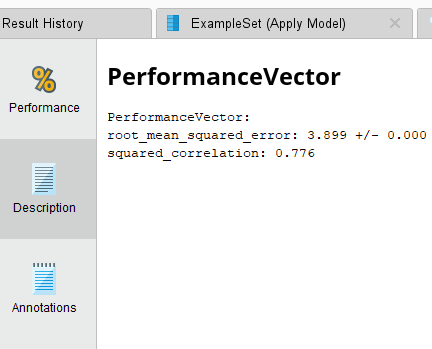
The RMSE is 3.607 if we apply the model on the training dataset

R square = 0.859, meaning: 85.9% of the variance in the label (MEDV) are explained by the attributes using the model on the training data set

This set of operators generate a model based on **the validation dataset (**the second split)



For this validation set, there are a few things we need to note:



The RMSE for the model being tested in validation is 3.899

R square is 0.776, meaning: 77.6% of the variance in the label (MEDV) are explained by the attributes using the model on the validation data set

Compare to the results for the training data set

RMSE(T) = 3.607 < RMSE(V) = 3.899

In general, we want RMSE to be small, and if RMSE for training (T) and Validation (V) are close that means the model we developed a good model.

If RMSE(V) < RMSE(T) that means the model does a really good job predicting new data

If RMSE(V) significantly bigger than RMSE(T) that means the model does a poor job (low predictive value) in predicting new data and it is likely that we over fit the data.

Our results suggest that the model does a good job given the RMSE are similar for both training and validation data set.